TIME-LAPSE OF PLANT MOVEMENT CLASSIFICATION

By:

AZRUL ZHAFRAN BIN AZREEZUL

(Matrix no: 125006)

Supervisor:

Dr. Loh Wei Ping

May 2018

This dissertation is submitted to
Universiti Sains Malaysia
As partial fulfillment of the requirement to graduate with honors degree in
BACHELOR OF ENGINEERING (MECHANICAL ENGINEERING)

School of Mechanical Engineering
Engineering Campus
Universiti Sains Malaysia

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed

(AZRUL ZHAFRAN BIN AZREEZUL)

Date

Statement 1

This journal is the result of my own investigation, except where otherwise stated. Other sources are acknowledged by giving explicit references. Bibliography/ references are appended.

Signed

(AZRUL ZHAFRAN BIN AZREEZUL)

Date

Statement 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for interlibrary loan, and for the title and summary to be made available outside organizations.

Signed

(AZRUL ZHAFRAN BIN AZREEZUL)

Date

ACKNOWLEDGEMENTS

In the name of Allah, the most beneficent, the most merciful. I wish like to express my deepest gratefulness:

To my supervisor, Dr Loh Wei Ping for her enormous guidance and advice during the research. It was great opportunity for me to do the research under her supervision with her experience in the data mining, applied biomechanics and mathematical modelling.

To my colleagues in the Biomotion Capture Laboratory who are also the final year students, Tang Jau Hoong and Ho Zhe Wei. It was a great pleasure to work along with them in the same supervision team. I would like to thank them all for the team work and support.

Last but not least, my best appreciation also goes to my parents and families for all the uncountable support and prayers. Their patience, understandings and care for providing the perfect environment for me to focus on my study are highly appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KLASIFIKASI LANGKAU MASA PERGERAKAN TUMBUHAN

ABSTRAK

Pergerakan tumbuhan biasanya diterokai dari perspektif tindak balas terhadap angin dan air. Apabila pergerakan tumbuhan terlalu lambat untuk diperhatikan dengan cepat, teknologi langkau masa menawarkan penyelesaiannya. Kajian terdahulu telah meneroka pergerakan tumbuhan dengan menggunakan pemodelan pokok dan simulasi botani. Konsep perlombongan data tumbuhan belum diguna pakai untuk mengkaji corak pergerakan. Walau bagaimanapun, tidak banyak kertas melaporkan corak pergerakan tumbuhan sebagai tindak balas terhadap gangguan luaran seperti angin, haba, cahaya dan air. Oleh itu, matlamat kajian ini adalah untuk mengklasifikasikan tindak balas tumbuhan dengan gangguan: angin dan air, membezakan kelas dengan jenis tumbuhan sebagai tindak balas terhadap gangguan angin dan air dan membandingkan corak pergerakan cabang ke arah angin atau air. Eksperimen dilakukan dengan menangkap gambar langkau masa pada lima jenis tumbuhan pasu terhadap tindak balas kepada angin dan air. Enam penanda ditempatkan di lokasi yang dikenal pasti dahan pokok (atas, tengah dan bawah) untuk membolehkan pengesanan gerakan. Video telah diterjemahkan ke dalam data berangka yang mana perubahan dalam pola tumbuhan akan dianalisis dengan kuantitatif menggunakan pendekatan perlombongan data. Peringkat yang terlibat termasuk (i) pra pengolahan data, (ii) klasifikasi (iii) penemuan pengetahuan. Teknik pra pengolahan data termasuk menormalkan, menyeragamkan dan mengeluarkan penjejak potensi dan nilai melampau. Pergerakan tumbuhan dikelompokkan ke dalam kelasnya: gangguan dan jenis tumbuhan berdasarkan keputusan pokok dan pengelap malas yang terbina dalam Weka. Analisis lanjut dilakukan untuk mengkaji jenis tumbuhan dan lokasi penanda yang mengakibatkan tidak dapat diklasifikasi. Hasil daripada kajian ini menunjukkan bahawa ketepatan klasifikasi 91.1745% diperolehi pada pengelas J48 untuk gangguan manakala 78.8992% untuk jenis tumbuhan.

TIME-LAPSE OF PLANT MOVEMENT CLASSIFICATION

ABSTRACT

Plant motions are commonly explored from perspectives of its responses to wind and water. As plant motion is too slow to be observed quickly, the time-lapse technology offers the solution. Previous studies have explored the movement of the plants by applying the tree modelling and botanical simulation. The data mining concept to study plant movements patterns is not adopted yet. However, not many papers reported the plant movement patterns in response to external perturbations such as wind, heat, light and water. Therefore, the goals of this study are to classify the plants responses by perturbation: wind and water, differentiate the classes by plant type in response to wind and water perturbations and compare the branches movement patterns towards wind or water. An experiment was conducted on time-lapse captures on five types of potted plants in response to wind and water. Six markers were placed on identified locations of tree branches (top, middle and bottom) to enable the motion tracking. The videos were translated into numeric data for which the changes in patterns of plants biomotion will be quantitatively analysed using data mining approach. Stages involved include (i) data preprocessing, (ii) classification (iii) knowledge discovery. Data preprocessing techniques include normalize, standardize and remove potential outlier and extreme value. The plants motion are grouped into its attribute classes: perturbation and plant type based on Decision Tree and Lazy classifiers built-in Weka tool. Further analysis was performed to examine the type of plant and location of markers that result in misclassifications. Findings from this study show that 91.1745% classification accuracies were retrieved on J48 classifier for perturbation while 78.8992% for type of plants.

CHAPTER 1

INTRODUCTION

1.0 OVERVIEW

Plants' growth and movements towards external perturbation may take weeks to observe. Throughout centuries, plant scientists put much efforts to study various aspects of botany ranging from the molecular (internal plant organs) to its physical movement responses. Plants movements are challenged by many disasters such as hurricane winds and monsoon rains and many other harsh mechanical perturbations that can threaten plant survival (Coutand, 2010). Plants' movement occurs over a wide range of sizes and time scales. For instance, the mimosa pudica responses quickly to touch but sunflower shows a slow response towards sunlight.

Previous studies have explored the movement of the plants by applying the tree modelling (Li *et al.*, 2011) and botanical simulation (Wang et al., 2017). Braam and Braam(2004) have considered the plant responses towards mechanical perturbations such as the touch stimulus. Some plant structures whereas use its mechanical instability concept to overcome hydraulic limit for faster movement (Forterre, 2013). Anten *et al.*, (2010) stated that plant responses to wind in can be variable depending on overall environmental conditions and plant characteristics. However, not many papers reported the plant movement patterns in response to external perturbations such as wind, heat, light and water. The plants motion patterns could be analyzed from data analysis perspectives. Apparently, the data mining concept to study plant movements patterns is not adopted so far. Data mining summarizes all attributes in order to identify interesting patterns (Gola *et al.*, 2018).

The study applies data mining concept to identify the plants responses to classify the plants responses by perturbation: wind and water, differentiate the classes by plant type in response to wind and water perturbations and compare the branches movement patterns towards wind or water.

This research focuses on potted plants time-lapse movements under wind and water factors. Time-lapse technology is used to display quick sequence of the plants movements.

The data collection involves experimental time-lapse video captures on five types of potted plants in responses to wind and water when simulated by standing fan (speed 3) and water shower respectively. At data analysis level, the study applies three stages of data mining: (i) data preprocessing, (ii) classification (iii) knowledge discovery. Collected data undergo qualitative data preprocessing that include data transformation to transform video-image-numeric and filtering by removing potential outlier and extreme value. Numeric attributing data extracted include perturbation, type of plant, point, x, y, dx, dy, distance, time and speed attributes. These data will be classified by perturbation and plant type classes. Knowledge discovery process involves the evaluation and interpreting of the patterns discover from classes and analyse the error based on branches location to identify the misclassified instances.

1.1 OBJECTIVE

The objectives of this study are to:

1. classify the plants responses by perturbation: wind and water
2. differentiate the classes by plant type in response to wind and water perturbations
3. compare the branches movement patterns towards wind or water

1.2 PROBLEM STATEMENT

Plants motion data analyses are often studied by using simulation but not yet analysed from time lapse perspectives. Past studies which explored plants simulation considers from perspectives of biomechanical model that match real trees (Wang et al., 2017). Li *et al.*, (2011) used a probabilistic approach to construct a dynamic 3D tree model from a 2D skeleton. At the same time, time-lapse video stores data which could be efficiently transformed into fruitful information and knowledge using the data mining concept. The combination of plants time-lapse captures with data mining analysis concept has not been applied in any studies. Data mining aspect enables user to extract useful information from large datasets. Data mining tools predict future trends and behaviours, thus patterns of the plants responses can be predicted. Existing studies have considered including markers to easily observe the motion of leaves and branches. However, there is no proper guideline on specific segment or location of the plants' branches to mark.

1.3 SCOPE OF PROJECT

This research adopts data mining applications in five types of potted plants' captured time-lapse motions towards wind and water perturbations. In the experiment domain: time-lapse captures of plant responses to two environmental stresses: wind and water simulated using table fan and shower respectively.

The attributes involved in this experiment include type of plant, perturbation, point, x, y, dx, dy, distance, time and speed. For this experiment, inconsistent water and wind effect is the main focus to reflect the actual environmental stresses.

The movement of the plant cannot be easily tracked by naked eyes. So the technology of time-lapse is applied to enable quick movements in short video play duration. Video data is extracted using Kinovea software. Data mining concept is applied at three stages: (i) data preprocessing, (ii) classification (iii) knowledge discovery in order to classify the plants responses by perturbation: wind and water and differentiate the classes by plant type in response to wind and water perturbations. The data mining analysis is aided by Weka tool.

CHAPTER 2

LITERATURE REVIEW

2.0 OVERVIEW

In this chapter, past studies on plants motions are presented along with the method of studies, followed by the plant responses toward external perturbations. Some parameter that such as effect leaf responses and the usage of markers to track the movement of plant are presented.

2.1 PLANT MOTION

Plants' motions are analysed in the context of the timescale of important physiological processes at the molecular and cellular level (Forterre, 2013). Plant motions are also considered from the perspectives of biomechanical model. Wang et al. (2017) produced animations that reasonably match the real trees by setting the mass density, stiffness and damping properties of branches and leaves. The authors had discovered that the ends of the branches can be seen clearly in videos as compared to leaves that almost rigid and cannot be easily tracked. Neubert et al. (2007) summarized the current tree modeling methods in three categories which is the rule-based generation, interactive modeling and image-based production.

Botanical simulation plays an important role in visual effects, games and virtual reality. Techniques applied involved the video featured point manual tracking in order to match real videos of trees. Stiffness (Young's modulus) and mass density were commonly assigned as these are the most important simulation parameters. More precise control could be achieved by exploiting the Poisson's ratio and other non-linear materials (Wang et al., 2017).

Li *et al.* (2011) showed a probabilistic approach for the programmed creation of tree models with persuading 3D appearance and movement. The only input was that the video of a moving tree that gives an initial dynamic tree model was utilised to produce new individual trees of the same type. The authors' approach combined the global and local constraints such as the shape of branches, the overall shape of the tree and physically plausible motions to construct a dynamic 3D tree model from a 2D skeleton.

2.2 TIME-LAPSE

The plants motions were hardly visible on naked eyes. Therefore, the technology of time-lapse were later introduced so that the underlying changes in the scene become quickly visible (Martin-Brualla et al., 2015). Campilho *et al.* (2006) used the time-lapse technology to analyze the stem-cell divisions in the Arabidopsis thaliana root meristem. Their experiment took 12 hours whereby the images were collected every 10 minutes of three different central z- planes.

The scientific and communicative value of time-lapse imagery enable data collection and ease visualization process. Time-lapse data sequences and linking time-lapse imagery with data visualization have made the ability to make data return alive and easy to understand (Brinley Buckley *et al.*, 2017).

2.3 PLANT RESPONSES TOWARD EXTERNAL PERTURBATIONS

Plants motions are always challenged by various disasters such as hurricane winds and monsoon rains and many other harsh mechanical perturbations that can threaten plant survival. Sensitive mechanisms have been developed by plants which can perceive and respond to stimuli (Coutand, 2010). Various signalling molecules and phytohormones have been involved in the touch responses as mentioned in Chehab et al. (2009). Mechanical perturbations such as wind and gravity are influenced morphogenesis (Braam and Braam, 2004). Plant experiences various types of environmental stresses such as drought, freezing, salinity and radiation which ultimately limits the total yield of a crop field (Chakraborty and Acharya, 2017).

Plants are subjected to a wide variety of abiotic and biotic stresses, which have been responsible for huge yield losses worldwide (Marques et al., 2017). Plants sense and react to numerous environmental signals that are assessed to competitively optimize acquisition of patchily distributed resources. Plant behaviour indicates the origins of plant intelligence that situated in complex communication (Trewavas, 2017).

Plants encounter a consistent changing environment, ranging from fast fluctuations of light and humidity caused by clouds, wind or rain, to larger diurnal and seasonal changes in temperature, light, rainfall and nutrient availability. In a few situations, plants have to deal with extreme conditions (Asensi-Fabado et al., 2017).

The main source of mechanical stimuli for plants are wind. When a plant is bent, its orientation in the gravity field is changed. On the other hand, when a plant is tilted, it bends under its own weight so that thigmomorphogenesis and gravitropism are not always easy to disentangle (Coutand, 2010).

Anten *et al.* (2010) showed that responses to mechanical stress and wind can be different and even in the opposite direction. Plant responses to wind can vary depending on overall environmental conditions and plant characteristics. Changes in environmental factors may influence multiple responses in plants, which will then predict the plant using data available.

Wind flow also exerts drag forces on plants and thus generate mechanical stresses. Plant responses to mechanical stress such as touch, rub and flexing can typically trigger inhibition of stem elongation, and increases in stem diameter and root allocation (Jaffe and Forbes, 1993).

Responses of plants towards wind will thus depend on air flow and mechanical stress effects which in turn depend on the overall environmental conditions as well as the characteristics of the plants themselves (Smith and Ennos, 2003). Plant responses may differ depending on the abiotic stress applied and the species. Smith and Ennos (2003) pointed out that the plant species plasticity will reflect the evolutionary history and current ecological conditions specific to each species.

Wind can also induce responses that are different or even opposite to those induced by pure mechanical stress (Henry and Thomas, 2002). Plants can respond to stress by adaptation, to temporal and spatial fluctuations in external stresses through adjustment in their shape and structure, allowing them to offset the impacts of the stress. Self-supporting terrestrial plants can be presented to a range of changing stresses, such as wind (Sellier and Fourcaud, 2005).

Plants can prevent mechanical failures under external forces such as wind and water flow by producing strong structures that oppose the large forces or by producing flexible structures that deflect and hence decrease the impact of forces (Read and Stokes, 2006).

2.4 EFFECTS OF LEAF MOVEMENT IN RESPONSE OF PLANT

The manner on how plants respond to wind load will rely on plant organ size and its mechanical properties such as leaf stiffness or stem flexibility. This may explain why the responses of different species to wind loading varies (Telewski, 1995).

Leaves can be repeatedly exposed to dynamic bending and twisting loads and abrasion with other foliage. It is often have large force yet are commonly light structures compared with stems and roots. (Niklas, 1992). Soft leaves can survive a moderate degree of mechanical stress. Even so, there is considerable variation in the structure and mechanics of leaves that is assumed to have adaptive significance in support and protection. However, the mechanical design of leaf laminae is poorly understood compared with that of petioles and stems (Niklas, 1999).

Leaves are probably most strongly influenced to wind effects on plants. The first reason is that the leaves are the primary organs of photosynthesis and transpiration, and the micro- climatic effects of wind affect them directly. Secondly, the leaves of most plants have large surface area to volume ratios (Niinemets and Fleck, 2002).

Plants are populated with leaves, which are widely studied in botany. Leaf elasticity and mass density mostly depend on the environment (Kirkham, 2005). A leaf with more water is stiffer and heavier. The leaves are often model as rigid.

2.5 USAGE OF MARKER TO TRACK THE MOVEMENT OF PLANT

Researchers used markers to track the movements of the plant. It is however hard to track when there are a lot of leaves and tangling branches. Xiong and Zhu (2001) used a series of marker genes (promoters) in different positions of the signal network.

High-throughput selective breeding techniques, such as marker assisted selection and high throughput phenotyping can accelerate the capacity to move the desirable traits into specific plant varieties by reducing the time it takes to identify progeny with the desired mutations (Collard and Mackill, 2008).

Johnová *et al.* (2016) reported that the exposure of plants to large amounts of different abiotic stresses such as osmotic and salt stress, leads to an increase in abscisic acid (ABA) content and to elevate expression of dehydration marker genes, whereas low levels of the similar types of stress affects a different combination of stress-responsive genes.

2.6 HIGHLIGHTS OF PREVIOUS WORKS

Plants are subjected to a wide variety of abiotic and biotic stresses. Leaves and branches play an importance role in tree modelling and simulation as it have large surface area to volume ratios (Niinemets and Fleck, 2002). It is importance to determine the parameter that affect plant motion. Past studies consider markers on leaves, this study extended the concept to perform markers on branches. The function of the markers are to make the observation of leaves and branches easier. Time-lapse were used so that the slowly changes in the scene become quickly visible. Plants motion data analyses are often studied by using simulation but not yet analysed from time-lapse perspectives. The combination of plants time-lapse captures with data mining analysis concept has not been applied in any studies.

CHAPTER 3

RESEARCH METHODOLOGY

3.0 OVERVIEW

This chapter presents the approaches adopted in studying the plants motion towards wind and water. Figure 3.1 shows the flowchart of the entire research methodology. Data collection experimental work to capture the time lapse of five types of plants' responses towards wind and water. A series of image frames was transformed from the time-lapse video captures. The images were translated into numeric data for which the changes in patterns of plants biomotion will be quantitatively analysed. The data mining analysis involves in this study under three main stages: (i) data preprocessing, (iii) classification and (iii) knowledge discovery.

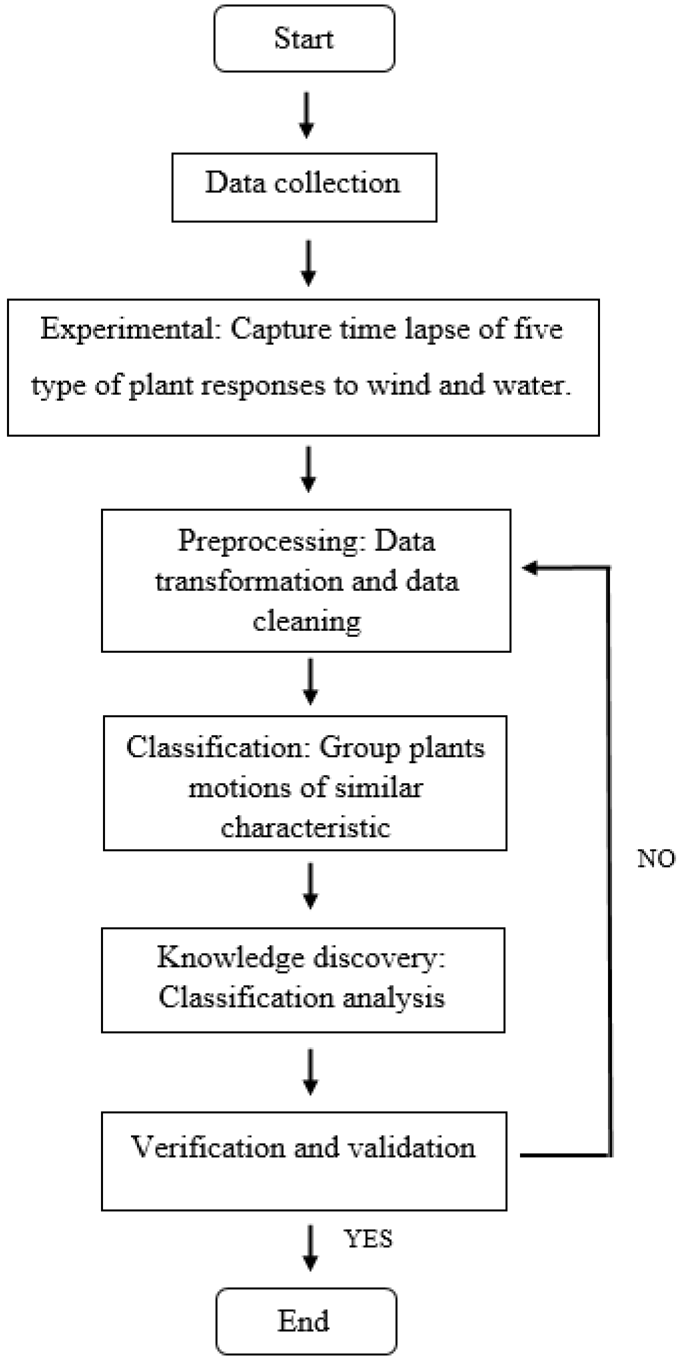

Figure 3.1: Flowchart of research methodology on response of plant toward wind and water

3.1 DATA COLLECTION

An experiment was carry out to collect data by capture time-lapse video of the plant. This experiment was conducted in Kangar, Perlis on 17 and 18 February 2018. The plants used were the potted plants that were available. There were five different types of plants with approximately same size, labelled as A, B, C, D and E in this study. Types of plants that were used in the experiment is shown in Figure 3.2.

Figure 3.2: Types of plants used in the experiment.

The experiment involves several sessions of time-lapse video captures of five types of plants (using Canon 650D installed with Magic Lantern firmware) while the plants were exposed to wind and water. The general entire experimental setup of the experiment is shown in Figure 3.3. The camera was set to shoot picture at every one second, beginning after 3 seconds and stop after 700 pictures were captured.

Figure 3.3: Experimental setup to capture plant motion.

The plants' motions towards wind and water were simulated using table fan and shower. The table fan were set to average speed of 3.5 m/s as this speed is sufficiently appropriate to show the response of plants. The water shower were applied from top approximately 3m in height.

3.2 DATA TRANSFORMATION

At data transformation level, the raw data which is time-lapse images are transformed into video followed by the numeric form. Each images were transformed into time-lapse using the Sony Vegas Pro 13 software. After the raw images is being transform into time-lapse video, the video is then run in Kinovea. Kinovea is a motion video analysis software to track the motion and enables visual comparison with reference frames. There were six markers labelled 1, 2, 3, 4, 5 and 6 at top, middle and bottom locations of the plants to study the motions at the particular segments (Figure 3.4). The markers were translated into coordinates data using Kinovea.

Figure 3.4: Location of six markers.

The step involves data cleaning to remove noisy and irrelevant data. These techniques eliminate the unwanted information in collected data. In this process, the data that is out of range is being eliminated.

The raw data were translated into 10 attributes and 42060 instances. The attributes include dx, dy, distance and speed calculated from the coordinates of x and y as shown in equations (3.1) – (3.4). These data were recorded in .csv format readable by Weka tool for further classification analysis as shown in Table 3.1.

$$dx = dx_2 - dx_1 \tag{3.1}$$

$$dy = dy_2 - dy_1 \tag{3.2}$$

$$Distance = \sqrt{dx^2 - dy^2} \tag{3.3}$$

$$Speed = \frac{Distance}{Time} \tag{3.4}$$

whereby

$x_1 = First\ point\ on\ x - coordinate$

$x_2 = Second\ point\ on\ x - coordinate$

$y_1 = First\ point\ o\ y - coordinate$

$y_2 = Second\ point\ on\ y - coordinate$

Table 3.1: Sample layout of raw data recorded in comma-separated format (.csv) format.

Pertubation	**Plant**	**Point**	**x**	**y**	**dx**	**dy**	**Distance**	**Time (s)**	**Speed**
Wind	A	1	0	0	0	0	0	0	0
Wind	A	1	0.06	0	0.06	0	0.06	1	0.06
Wind	A	1	0.29	0.06	0.23	0.06	0.237697	1	0.237697
Wind	A	1	0.06	0.06	-0.23	0	0.23	1	0.23
Wind	A	1	-0.35	0	-0.41	-0.06	0.414367	1	0.414367
Wind	A	1	-0.35	0.06	0	0.06	0.06	1	0.06
Wind	A	1	-0.17	0.06	0.18	0	0.18	1	0.18
Wind	A	1	0.52	0.06	0.69	0	0.69	1	0.69
Wind	A	1	-0.29	0	-0.81	-0.06	0.812219	1	0.812219
Wind	A	1	-0.29	-0.06	0	-0.06	0.06	1	0.06

3.3 DATA MINING

The recorded numeric raw data undergo data mining analysis: data preprocessing, classification and knowledge deployment using Weka software.

3.3.1 PREPROCESSING

The datasets was normalized and standardized to check the accuracy thus can compare with raw data. The raw data is normalized by rescaling one or more attributes to the range of 0 to 1 (Ian H. Witten and Frank, 2005). This means that the smallest value in dataset is 0 while the largest is 1. The purpose of normalizing is eliminating redundancy and inconsistent dependency. Data standardization is the process of rescaling all numeric attribute to have a mean value of 0 and a standard deviation of 1.

Data classification analysis further refined on qualitative data enhancement by removing potential outlier and extreme value. The potential outliers and extreme values were screened based on equations (3.5) to (3.8) respectively.

$$Q3 + OF \times IQR < x \leq Q3 + EVF \times IQR \tag{3.5}$$

$$or$$

$$Q1 - EVF \times IQR \leq x < Q1 - OF \times IQR \tag{3.6}$$

$$x > Q3 + EVF \times IQR \tag{3.7}$$

$$or$$

$$x < Q1 - EVF \times IQR \tag{3.8}$$

whereby

$Q1 \ = \ 25\% \ quartile$

$Q3 \ = \ 75\% \ quartile$

$IQR \ = \ Interquartile\ Range, difference\ between\ Q1\ and\ Q3$

$OF \ = \ Outlier\ Factor$

$EVF \ = \ Extreme\ Value\ Factor$

The identified outliers, extreme value, time, dx, dy and distance were removed. This is due to contribution to error thus effect the percentage of correctly classified. Having outliers, extreme value, time, dx, dy and distance will result in inaccurate or bias results in further classification analysis.

3.3.2 **CLASSIFICATION**

Data classification were run for all classifiers built-in Weka tool: bayes, functions, lazy, meta, misc, trees and rules. All algorithm in these classifier were performed to determine percentage of correctly classified. Best two classfier were choose based one it accuracy and consistency which is Decision Tree and Lazy. Decision tree is a flow-chart like structure, where each branch represents the outcome of a test and each leaf represents a class label. There are seven algorithm available in Decision Tree classifier which are DecisionStump, HoeffdingTree, J48, LMT, RandomForest, RandomTree and REPTree. Lazy classifier calculate the distances between the testing example and all of the training data in order to identify its nearest neighbors through all dataset. There are three algorithm in Lazy classifier which are IBK, KStar and LWL. There were two algorithms which is J48 from Decision Tree and IBK from Lazy classifier employed in this study.

The plants motion having similar patterns are grouped into its attribute classes: perturbation and plant. There are only two classes for perturbation: wind and water. This research also consider type of plant as class. There are five classes which is Plant A, Plant B, Plant C, Plant D and Plant E. The algorithm were tested using 10 folds cross-validation using default Weka setting. This means that the dataset is split into 10 equally-sized folds.

3.3.3 **KNOWLEDGE DEPLOYMENT**

A further perturbation classification analysis on J48 and IBK were performed on distinctive plant type (single plant at a time) and single perturbation (either water or wind). The confusion matrix was used to evaluate the classifier quality. For this dataset, two classes is obtained. In the two class case, results can be summarized in the form of a 2x2 matrix in which the diagonal elements represent correct classification and other elements represent error (Witten and Frank, 2005). General form of confusion matrix shown in Figure 3.5. The confusion matrix contains information about actual and predicted classifications done by a classification system. The correct number of instances, wrong number of instances and error were calculated based on equations (3.9) to (3.11) respectively.

$$\begin{array}{cc|c} a & b & \\ \hline s & t & a \\ u & v & b \end{array}$$

Figure 3.5: General form of confusion matrix.

$$Correct\ number\ of\ instances = s + v \tag{3.9}$$

$$Wrong\ number\ of\ instances = t + u \tag{3.10}$$

$$Error = \frac{t + u}{s + t + u + v} \times 100\% \tag{3.11}$$

Error analysis was performed to examine which part of plant markers cause the misclassifications. There were six markers with different locations labeled: top, middle and bottom. Three points: 1-3 were marked at top, two points: 4-5 marked at the middle and one point: 6 was marked at bottom of the plant branch. Misclassifications shown in matrices were investigated by these marker points, marker location and type of plant.

CHAPTER 4

RESULTS AND DISCUSSION

4.0 OVERVIEW

This chapter presents the results of the study specifically from data classification analysis. Based on evaluations on all embedded classifiers in Weka, the Decision Tree and Lazy classifiers were found more appropriate in terms of classification accuracies and reliability from the nature of the study data. The classification performances were evaluated from percentage of accurately classified instances and number of instances in error analysis.

4.1 DECISION TREE AND LAZY CLASSIFIER

The raw study dataset contain 10 attributes (perturbation, type of plant, point, x, y, dx, dy, distance, time and speed) and 42060 instances. On data preprocessing analysis, the dataset were normalized and standardized. Normalization resulted in attributes to have range of 0 to 1 while standardization resulted attribute to have a mean value of 0 and a standard deviation of 1. Meanwhile, there were 5725 data outliers and 22788 extreme values identified and being discarded from the raw data. There were two attributes predefined as the data class: perturbation (wind and water) and plant type (A, B, C, D, E). Using seven algorithms of Decision Tree and three algorithms of Lazy classifier, the percentage classification accuracy were investigated on the raw and preprocessed data: normalized, standardized and outliers.

Table 4.1 and 4.2 shows the results of percentage of correctly classified data by perturbation class when data were classified into two classes by perturbation and five classes by plant type. Based on the results from Table 4.1 and 4.2, two best performed algorithm for all type of dataset are J48 and IBK showed 91.1745% and 90.6871% accuracy for perturbation while 78.8992% and 78.2192% for plant type. Meanwhile, other algorithms like DecisionStump merely show accuracy around 63.6053%-75.9651% for perturbation while 32.1541%-41.0284% for plant type.

When the datasets was normalized and standardized, the classification accuracies slightly change within 78.2216%-78.2382% by performed IBK algorithm for plant. However, some algorithms like DecisionStump, J48, LMT and LWL remain unchanged.

Owing to not much difference observed, classification analysis further refined on qualitative data enhancement by removing potential outlier and extreme value. Nevertheless, the percentage of correctly classified mostly decreased by 0-1.3909% as compared to the original raw data for all algorithm.

At another level classification enhancement analysis, the attributes: time, dx, dy and distance filtration was considered. Table 4.1 and 4.2 shows that the classification accuracy become weaker. This is because of some attribute are the compound measures of other basic attributes. For instance the value of *dx* and *dy* is calculated by using position of *x* and *y* while distance is calculated using value of *dx* and *dy*. Therefore, removing compound attribute made no difference to removing basic attribute. Meanwhile, the time attribute does not effect on the accuracy because it remained constant with time-step 1 second throughout the experiments.

Table 4.1 Percentage of correctly classified data by perturbation class

Classifier	Type	RAW (%)	Filter		Remove					
			Normalize (%)	Standardize (%)	Outlier (%)	Extreme (%)	Time (%)	Dx dy (%)	Time dx dy (%)	Time dx dy distance (%)
Trees	DecisionStump	68.2311	68.2311	68.2311	63.6053	75.9651	68.2311	68.2311	68.2311	68.2311
	HoeffdingTree	84.0656	84.4532	84.2725	82.7054	82.9182	84.2297	84.8764	84.7979	84.5388
	J48	91.1745	91.1745	91.1745	89.8473	85.9641	91.1793	91.0319	91.0366	91.0342
	LMT	91.2506	91.2601	91.2577	89.9161	86.1094	91.3053	91.1317	91.1959	91.1888
	RandomForest	91.5002	91.5193	91.5169	90.3647	86.1665	91.5668	91.2791	91.3647	91.436
	RandomTree	91.0747	91.0865	91.1484	89.8418	86.1249	91.1698	91.0247	91.0152	91.0794
	REPTree	90.7941	90.7965	90.7917	89.7041	85.9641	90.8036	90.6633	90.6728	90.6728
Lazy	IBK	90.6871	90.6871	90.6871	89.4702	86.1146	90.7394	90.4969	90.5492	90.5516
	KStar	89.3058	89.3058	89.3058	87.9015	84.2206	89.3058	89.2011	89.2011	89.1774
	LWL	78.1526	78.1526	78.1526	75.8938	75.9651	78.1503	78.4855	78.533	79.0442

Table 4.2: Percentage of correctly classified data by plant class

Classifier	Type	RAW (%)	Filter		Remove					
			Normalize (%)	Standardize (%)	Outlier (%)	Extreme (%)	Time (%)	Dx dy (%)	Time dx dy (%)	Time dx dy distance (%)
Trees	DecisionStump	32.1541	32.1541	32.1541	34.5837	41.0284	32.1541	32.1541	32.1541	32.1541
	HoeffdingTree	50.7751	52.166	51.8759	51.2591	48.3966	51.1246	49.0609	48.3928	49.7575
	J48	78.8992	78.8992	78.8992	76.2158	65.5044	78.9087	77.3395	77.2967	77.2943
	LMT	79.0038	79.0014	78.9967	76.2295	65.6963	79.0418	77.2563	77.2706	77.2682
	RandomForest	79.5078	79.5078	79.4579	76.6974	65.8676	79.5364	77.4536	77.5487	77.5392
	RandomTree	78.5164	78.4736	78.6091	75.7617	65.6652	78.7565	76.826	76.902	77.0162
	REPTree	78.1479	78.1526	78.1479	75.4507	65.2345	78.1431	76.776	76.7808	76.7808
Lazy	IBK	78.2192	78.2216	78.2382	75.6571	65.66	78.2454	75.825	75.8512	75.8726
	KStar	74.9453	74.9453	78.2382	72.1068	62.5571	74.9453	72.4727	72.4727	72.0946
	LWL	37.5178	37.5178	37.5178	40.3798	44.6088	37.5392	37.7485	37.7722	37.7199

Figure 4.1 shows the raw data perturbation classification resulted from J48 classifier indicating 38348 correctly classified instances with 91.1745% accuracy. There are 3712 instances that incorrectly classified with 8.8255% error.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       38348               91.1745 %
Incorrectly Classified Instances      3712                8.8255 %
Kappa statistic                          0.8235
Mean absolute error                      0.1265
Root mean squared error                  0.2567
Relative absolute error                 25.308  %
Root relative squared error             51.3467 %
Total Number of Instances            42060

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.855    0.031    0.965      0.855    0.906      0.829    0.965     0.969     Wind
                 0.969    0.145    0.870      0.969    0.917      0.829    0.965     0.957     Water
Weighted Avg.    0.912    0.088    0.917      0.912    0.911      0.829    0.965     0.963

=== Confusion Matrix ===

     a     b   <-- classified as
 17976  3054 |     a = Wind
   658 20372 |     b = Water
```

Figure 4.1: Perturbation classification results executed from J48 algorithm on raw data.

Based on confusion matrix inspection, , there were 3054 instances that incorrectly predict as water while its actual class is wind and 658 instances that incorrectly predict as wind while its actual class is water. Meanwhile on IBK algorithm, there were 3203 instances incorrectly classified as water and 714 instances incorrectly classified as wind (Figure 4.2). Both algorithms agreed that more wind perturbation on plants was mistaken as water.

```
     a     b   <-- classified as
 17827  3203 |     a = Wind
   714 20316 |     b = Water
```

Figure 4.2: Confusion matrix generated from perturbation classification using IBK algorithm on raw data.

Figure 4.3 shows the confusion matrices developed on plant type classification using J48 and IBK algorithm.

```
    a    b    c    d    e   <-- classified as
 5244  213  241 2198  516 |    a = A
  476 7039   96  742   59 |    b = B
  389  320 5860 1727  116 |    c = C
  419    1  108 7883    1 |    d = D
  267   77  159  750 7159 |    e = E
```

(a)

```
    a    b    c    d    e   <-- classified as
 5167  238  278 2147  582 |    a = A
  477 7091   74  717   53 |    b = B
  409  323 5840 1663  177 |    c = C
  432    2  196 7779    3 |    d = D
  336   85  236  733 7022 |    e = E
```

(b)

Figure 4.3: Confusion matrix of plant type classification using (a) J48 algorithm and (b) IBK

Table 4.3 summarizes the number of instances incorrectly grouped as different plant type observed from Figure 4.3 on two algorithms, J48 and IBK.

Table 4.3: Number of instance error that incorrectly classified by plant type.

Plant	Error	
	J48	IBK
A	3168	3245
B	1373	1321
C	2552	2572
D	529	633
E	1253	1390

On J48 algorithm, there were 5244 out of 8412 instances of Plant A correctly classified but incorrectly classified 213 as Plant B, 241 as Plant C, 2198 as Plant D and 516 as Plant E. On IBK algorithm, there were 7779 out of 8412 instances of Plant D correctly classified but 432 incorrectly classified as Plant A, 2 as Plant B, 196 as Plant C and 3 as Plant E. Based on the table 4,5, Plant A contributes most error while Plant D contribute least error. This could be due to the size of the branch of Plant A is much smaller than Plant D (Figure 4.3). Thus the response toward wind and water is clearer in Plant D.

(a) (b)

Figure 4.4: Comparison between (a) Plant A and (b) Plant D

4.2 PERTURBATION ANALYSIS BY PLANT TYPE AND VICE VERSA

Having identified the classification accuracies by perturbation (wind and water) and by plant type (A-E), it is interesting to know any type of plant well distinguish the perturbation impact. Therefore, a further perturbation classification analysis on J48 and IBK were performed on distinctive plant type (single plant at a time). The results were shown by confusion matrices in Table 4.4 and summarized in Figure 4.5. From the confusion matrices, the wind perturbation contributes more error as compared to water. This is observed in Plant D.

Table 4.4: Perturbation classification on distinctive plant using J48 and IBK algorithm.

<table>
<tr><th rowspan="2">Plant</th><th colspan="2">J48</th><th colspan="2">IBK</th></tr>
<tr><th>Correctly Classified</th><th>Confusion Matrix</th><th>Correctly Classified</th><th>Confusion Matrix</th></tr>
<tr><td>A</td><td>86.7927 %</td><td>a b <-- classified as
3387 819 | a = Wind
292 3914 | b = Water</td><td>87.1374 %</td><td>a b <-- classified as
3398 808 | a = Wind
274 3932 | b = Water</td></tr>
<tr><td>B</td><td>99.8692 %</td><td>a b <-- classified as
4197 9 | a = Wind
2 4204 | b = Water</td><td>99.8811 %</td><td>a b <-- classified as
4202 4 | a = Wind
6 4200 | b = Water</td></tr>
<tr><td>C</td><td>88.1479 %</td><td>a b <-- classified as
3412 794 | a = Wind
203 4003 | b = Water</td><td>87.9339 %</td><td>a b <-- classified as
3399 807 | a = Wind
208 3998 | b = Water</td></tr>
<tr><td>D</td><td>86.2696 %</td><td>a b <-- classified as
3057 1149 | a = Wind
6 4200 | b = Water</td><td>86.3885 %</td><td>a b <-- classified as
3074 1132 | a = Wind
13 4193 | b = Water</td></tr>
<tr><td>E</td><td>95.0309 %</td><td>a b <-- classified as
3912 294 | a = Wind
124 4082 | b = Water</td><td>93.3547 %</td><td>a b <-- classified as
3822 384 | a = Wind
175 4031 | b = Water</td></tr>
</table>